Analisi di segmentazione con una variabile dipendente trasformata in *logit*

Luigi Fabbris, Maria Cristiana Martini[1]
Dipartimento di Scienze Statistiche, Università di Padova

Riassunto: L'analisi di segmentazione è un metodo di analisi statistica multivariata mirato alla partizione progressiva di un campione di unità caratterizzate da una variabile dipendente e da un insieme di variabili predittive osservate su una qualsiasi scala. Se la variabile dipendente è dicotomica, l'analisi di segmentazione convenzionale è problematica, per cui si propone di trasformare la proporzione di casi che possiedono l'attributo rappresentato dalla variabile dipendente nel *logit* della frequenza. La partizione del campione è valutata, ad ogni passo dell'analisi, con riferimento alla massima differenza tra i *logit* delle proporzioni nei sottocampioni che si formano con la scissione. La segmentazione è valutata statisticamente anche in relazione alle opzioni *look-ahead* per la ricerca di interazioni, relazione monotòna tra la variabile dipendente e un predittore ordinale, penale per ricerca di alberi a struttura
Parole chiave: Analisi di segmentazione; Trasformazione *logit*; Ricerca di interazioni; Relazione monotòna; Alberi di regressione; Simmetria dell'albero.

1. L'analisi di segmentazione binaria di campioni

L'analisi di segmentazione è un metodo di analisi statistica multivariata mirato alla partizione progressiva di un campione di unità statistiche caratterizzate da una variabile dipendente e da una pluralità di variabili predittive osservate su qualsiasi scala. La funzione obiettivo per la scissione progressiva del campione è determinata in ragione delle caratteristiche tecniche della variabile dipendente.

[1] Il presente lavoro è stato finanziato nell'ambito del PRIN "La ricerca di determinanti del rischio mediante analisi di segmentazione di campioni". Coordinatore nazionale e dell'Unità di Padova è Luigi Fabbris. La riflessione metodologica che ha portato alla formulazione della proposta illustrata nella nota è stata svolta congiuntamente dagli autori. Il Par. 3 del testo è stato redatto da M.C. Martini, i restanti paragrafi da L. Fabbris. Gli autori desiderano ringraziare il dott. Carlo Schievano e i referee per alcuni preziosi consigli in fase di stesura della nota.

Si denoti con Y la variabile dipendente e con $\mathbf{x}'=\{X_i, i=1, \ldots, q\}$ il vettore delle q variabili predittive osservate su ciascuna delle n unità statistiche. Se Y è dicotomica, si può applicare una tecnica scissoria basata sulla massimizzazione della distanza tra le frequenze (Sonquist *et al.*, 1973). Tuttavia, la procedura si comporta in modo indesiderato, generando molti gruppi con frequenze disperse, quando la frazione di casi che possiedono la caratteristica Y nel campione è piccola, per esempio si tratta della frequenza di una malattia (Fielding, 1977).

In questa nota si propone, a fini analitici, la sostituzione della variabile dipendente dicotomica con il *logit* della probabilità condizionata di Y, cioè con il logaritmo naturale del rapporto tra tale probabilità e il suo complemento a uno:

$$logit\big(\pi(Y \mid \mathbf{x})\big) = ln\left(\frac{\pi(Y \mid \mathbf{x})}{1-\pi(Y \mid \mathbf{x})}\right), \tag{1}$$

dove $\pi(Y|\mathbf{x})$ denota il valore di Y condizionato da un insieme di predittori e *ln* il logaritmo naturale dell'argomento entro parentesi. La probabilità $\pi(Y|\mathbf{x})=P(Y=1|\mathbf{x})$ varia tra 0 e 1. D'ora in avanti, per semplicità, scriveremo π invece di $\pi(Y|\mathbf{x})$.

La procedura di segmentazione ripartisce il campione con una sequenza di suddivisioni dicotomiche, o binarie, dei gruppi che via via si formano nell'analisi. La prima suddivisione dicotomica riguarderà il gruppo iniziale di n unità, il quale sarà suddiviso in due sottogruppi aggregando opportunamente le modalità del più efficace tra i predittori candidati, la seconda sarà la migliore tra le due migliori suddivisioni dei gruppi appena formati, sempre in base all'aggregazione delle modalità dei predittori candidati più efficaci, la terza sarà la migliore tra quelle possibili sui tre gruppi formati, e così di seguito finché si verifica una regola di arresto del processo (per maggiori dettagli, si può consultare Fabbris, 1997: cap. 9). I gruppi così formati generano un albero a struttura gerarchica, detto *dendrogramma*.

Oltre che binaria, la suddivisione può essere ternaria, ossia con tripartizione del campione, o addirittura politomica. Criteri di segmentazione binaria e ternaria sono presenti nei testi che accompagnano programmi *software*. Tra i più noti si possono citare: AID (Sonquist *et al.*, 1973), ELISEE (Cellard *et al.*, 1967), THAID (Morgan e Messenger, 1973), CART (Breiman *et al.*, 1984), CHAID (Kass, 1980), C4.5 (Quinlan, 1993), TREE (Venables e Ripley, 1994), UNAIDED (Capiluppi e Fabbris, 1998; Capiluppi *et al.*, 1999). Le segmentazioni binarie o ternarie soddisfano la generalità delle esigenze di ricerca.

Qualificano l'analisi di segmentazione le seguenti opzioni:

- la forzatura nelle prime fasi dell'analisi di una o più variabili che possono essere di disturbo – in quanto connesse sia alla variabile dipendente che a quella predittiva – nella valutazione della relazione di dipendenza tra queste due;
- l'ordinamento dei predittori secondo un ordine causale predefinito. In questo modo, non solo si evitano selezioni casuali (Flack e Chang, 1987), ma la ricerca di

predittori prima nella categoria che comprende quelli causalmente più lontani, poi tra quelli successivi nell'ordinamento imposto, implica che il processo di selezione segua la logica della causalità;
- la suddivisione dei campioni sulla base di interazioni tra variabili, oltre che sulla base di singole variabili, detta procedura *look-ahead*;
- per predittori ordinali, il considerare ammissibili le combinazioni libere delle modalità, oppure il mantenere l'ordinalità delle modalità osservate (quest'ultima opzione è detta combinazione "monotòna" delle modalità, l'alternativa è detta combinazione "libera");
- la ricerca di soluzioni sub-ottimali che, però, mantengono la simmetria dell'albero, ossia la ricerca di suddivisioni sulla base di poche variabili, qualificate dalle stesse modalità, al fine di procedere con analisi di classificazione;
- altre opportunità di analisi, quali il cosiddetto *pruning*, o sfoltimento, dell'albero dopo averlo espanso fino ad un limite, la fusione di sottogruppi simili ad un certo livello dell'analisi, facendo così perdere la gerarchia delle suddivisioni, e altre possibilità di manipolazione del dendrogramma.

Nel seguito della nota, si esaminano i criteri per l'ottimizzazione del processo di aggregazione delle modalità del predittore candidato (Par. 2), la valenza statistica della suddivisione del campione fondata sui criteri ottimi (Par. 3), i criteri di ripartizione del campione basati sulle interazioni tra predittori candidati (Par. 4), la logica della penale per simmetria (Par. 5). La nota si conclude (Par. 6) con una valutazione critica della proposta e con l'indicazione di direzioni per lo sviluppo ulteriore del processo di segmentazione ideato.

2. Segmentazione binaria del campione con riferimento ad una variabile dipendente binaria trasformata in *logit*

Si considerino la variabile dipendente dicotomica Y e la variabile predittiva $X=\{X_i, i=1, \ldots, K\}$, le cui modalità sono in numero di K. Nel modello asimmetrico ($Y \leftarrow X$), X rappresenta una delle q variabili predittive candidate per la partizione del campione in funzione della Y.

La combinazione delle modalità di X che si prende in esame per la partizione del campione è quella che rende massimo lo scarto tra i *logit* delle frequenze relative di Y. I sottogruppi per i quali si calcolano le frequenze di Y si ottengono combinando in due insiemi le modalità della variabile X.

Analiticamente, la suddivisione ottima corrisponde alla partizione del campione che rende massimo il valore di δ [*logit*($\pi(Y|X_1, X_0)$)]:

$$\left[logit(\hat{\pi}(\mathrm{y} \mid \mathrm{X}_1)) - logit(\hat{\pi}(\mathrm{y} \mid \mathrm{X}_0))\right] = \mathrm{Max}\left[ln\ \frac{\hat{\pi}_1(1-\hat{\pi}_0)}{(1-\hat{\pi}_1)\hat{\pi}_0} \right] \quad (2)$$

$$= \mathrm{Max}\left[ln\ \frac{n(y \mid X_1)\ [n - n(y \mid X_0)]}{[n - n(y \mid X_1)]\ n(y \mid X_0)} \right], \quad (3)$$

dove: X_1 e X_0 denotano due insiemi complementari di modalità di X;

$\hat{\pi}(Y \mid X_1) = p(Y|X_1)=p(Y_1)>p(Y_0)$ denota la frequenza relativa di Y condizionatamente alla categoria X_1.

L'argomento del logaritmo delle formule (2) e (3), detto *rapporto crociato* (in inglese *odds ratio*[2]), misura quanto la Y è determinata da X_1 in rapporto a X_0.

Il criterio $\delta[logit(\pi(\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_0))]$ diventa infinito ogni volta che le frequenze di y si concentrano su uno solo dei due insiemi di categorie, lasciando frequenze nulle nell'insieme complementare[3]. Siccome questa eventualità è proprio ciò che si cerca con la segmentazione, si può adottare una o più delle seguenti opzioni:

- aggiungere un valore alle frequenze della formula (3) sufficientemente piccolo da non alterare l'ordine di significatività delle suddivisioni;
- applicare un test per la verifica della significatività statistica parallelamente all'analisi degli scarti tra i *logit* (cfr. Par. 3);
- in ogni caso, per dare consistenza all'analisi, bisogna evitare che le frequenze n_i di X_i siano inferiori ad uno standard prefissato, ossia bisogna impedire che l'analisi delle frequenze di y sia svolta entro sottocampioni troppo esigui.

Tabella 1. *Ordinamento delle modalità di* X *secondo la proporzione* p_i *di unità che possiedono l'attributo Y*

Modalità di X	*Proporzione di* Y	
X_1	p_1	Max $p(Y\|X)$
:	:	:
X_i	p_i	..
:	:	:
X_K	p_K	Min $p(Y\|X)$

[2] In inglese *odds* significa rapporto di probabilità. *Odds ratio* (in italiano: rapporto crociato) è il rapporto tra due rapporti tra probabilità complementari (Hosmer e Lemeshow, 1989).

[3] Si tratta di un problema comune anche ad altri approcci; per esempio, l'indice basato sull'entropia ha un problema analogo, dato che contiene nella sua formula il logaritmo di una quantità che può avere valore zero.

Tabella 2. *Tentativi di suddivisione binaria del campione in base all'ordinamento delle modalità di* x *secondo la proporzione di unità che possiedono il carattere* y

		I tentativo	II tentativo	….	*k*-1 tentativ
X_1	p_1	p_1	p_1	..	p_1
X_2	p_2	p_2	p_2	:	:
X_3	p_3	:	p_3	:	:
:	:	:	:	:	p_{K-1}
X_K	p_K	p_K	p_K	..	p_K

La procedura di combinazione delle modalità del predittore candidato *x* che rende ottima la suddivisione del campione è la seguente:

a. ordinare in senso decrescente le *K* modalità del predittore *x* secondo la proporzione p_i (i=1, …, K) di unità che possiedono l'attributo *Y* (Tab. 1);
b. valutare statisticamente i *K*-1 tentativi di partizione che si realizzano suddividendo in forma binaria, volta a volta, il campione nelle (*K*-1) coppie di sottocampioni che si formano scendendo da p_1 verso p_K (Tab. 2) e calcolando per ciascun tentativo il valore di δ [*logit*(π(Y|X_1, X_0))];
c. individuare la migliore partizione in base al massimo valore di δ[*logit*(π(Y|X_1, X_0))] tra i *K*-1 *logit*:

$$\Delta[logit(\pi(Y|X_1, X_0))] = \text{Max } \{\delta[logit(\pi(Y|X_1, X_0))] \}; \quad (4)$$

d. verificare la significatività statistica di Δ[*logit*(π(Y|X_1, X_0))] e, se la suddivisione è significativa, realizzare la partizione e diramare l'albero.

La suddivisione così definita è la migliore in assoluto, anche delle combinazioni di modalità di *X* ignorate nella procedura, ossia delle restanti modalità di *X* non ordinate secondo la proporzione del carattere *Y* (cfr. anche Fisher, 1958).

La statistica Δ[*logit*(π(Y|X_1, X_0))] deriva da tre livelli di confronto:

1. fra partizioni alternative delle modalità per scegliere la migliore suddivisione, all'interno di ciascun predittore, per ogni nodo candidato alla suddivisione;
2. fra le migliori suddivisioni di ogni predittore, entro ciascun nodo, allo scopo di individuare la variabile che garantisce la migliore partizione del nodo stesso;
3. fra le migliori partizioni di ogni nodo, al fine di individuare quello per cui si ottiene la migliore partizione in assoluto.

La procedura è banale sul piano pratico se si impone la monotonicità della relazione tra il *logit* del valore atteso di *Y* e la variabile candidata alla suddivisione. Infatti, le modalità rimangono nella stessa successione nella quale sono state osservate.

3. Valutazione statistica della suddivisione

Una filosofia comune a diversi criteri di segmentazione è legata alla definizione di una misura di disomogeneità dell'albero "di classificazione", o "di regressione", e alla ricerca della suddivisione che riduce massimamente tale disomogeneità. La misura di disomogeneità può essere definita in molti modi diversi, per esempio in funzione della varianza, dell'entropia o della mutabilità, secondo la natura della variabile considerata come dipendente e della tecnica proposta[4].

Se guardiamo alla segmentazione di campioni come ad un metodo per la ricerca di determinanti del rischio, l'interpretazione della statistica proposta nel Par. 2 come criterio per la selezione della suddivisione successiva è certamente più immediata, trattandosi di una funzione della approssimazione del rischio relativo, nonché del coefficiente relativo all'effetto principale o di interazione corrispondente alla suddivisione in esame in una regressione logistica.

Nel seguito, si valutano le caratteristiche del criterio proposto al confronto con gli approcci basati sulla varianza della proporzione $p(Y=1)$ e sul test χ^2, con riguardo alle situazioni di frequenze marginali basse o squilibrate, oppure di frequenze nulle o molto piccole all'interno della tabella di relazione tra X e Y (Tab. 3).

Tabella 3. *Tabella di frequenze per l'analisi dell'effetto della segmentazione (le frequenze in grassetto sono parametri dell'esperimento)*

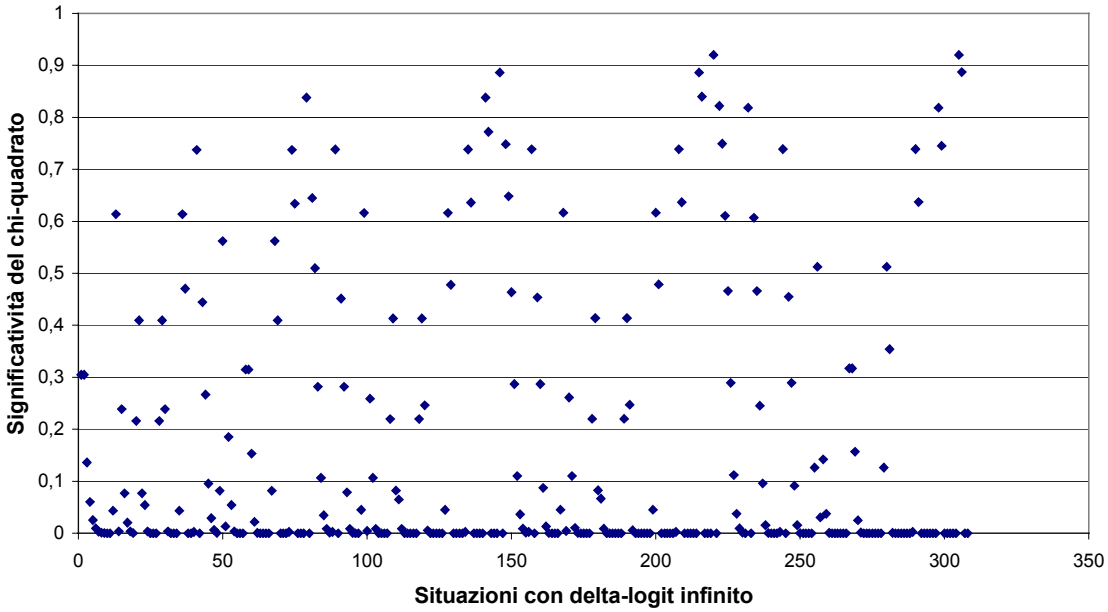
		Dipendente		
		$Y=1$	$Y=0$	
Predittore	$X=1$	***a***	b	***a+b***
	$X=0$	c	d	$c+d$
		a+c	$b+d$	***n***

A tale scopo, si calcolano alcune statistiche criterio su un insieme di 856 diverse tabelle tetracoriche, corrispondenti a suddivisioni alternative di nodi diversi dell'albero di regressione. Le situazioni sperimentali sono generate controllando i seguenti parametri:

[4] Nel caso in esame, in cui la variabile dipendente è dicotomica, il metodo basato sull'indice di mutabilità di Gini e quello basato sulla varianza coincidono. Breiman *et al.* (1984) osservano che, per una variabile dipendente dicotomica, il criterio basato sull'entropia ha proprietà formali analoghe all'indice di mutabilità di Gini e a quello basato sulla varianza; esso fornisce risultati simili, ma presenta una maggiore complessità di calcolo. Quindi, se si cercano soluzioni semplici, sono preferibili gli altri due al criterio basato sull'entropia. Ancora diversa è la logica sottostante al metodo CHAID proposto da Kass (1980) per variabili dipendenti nominali, che consiste nel compiere suddivisioni significative sulla base del test χ^2.

1. *la numerosità* (n) *del nodo considerato per la suddivisione*. La numerosità totale del nodo è stata fatta variare da 20 a 1000;
2. *il numero* (a+c) *di unità che possiedono l'attributo* Y *nel nodo*. Il numero di unità che possiedono l'attributo è stato fatto variare tra un minimo di un caso e un massimo del 50% della popolazione, dato che, quando $p(Y=1)>0,5$, il valore delle statistiche è simmetrico;
3. *la numerosità* (a+b) *della classe* X=1 *formata dalla suddivisione della variabile predittiva* X. Per la numerosità della classe $X=1$ si sono escluse situazioni che generassero classi con meno di 10 unità;
4. *il numero* a *di unità della classe* X_1 *che possiedono l'attributo* Y=1. Si sono considerate situazioni in cui nessuno, uno o due delle unità della classe X_1 possiedono l'attributo $Y=1$.

Figura 1: Valori del chi-quadrato per valori infiniti di delta-logit



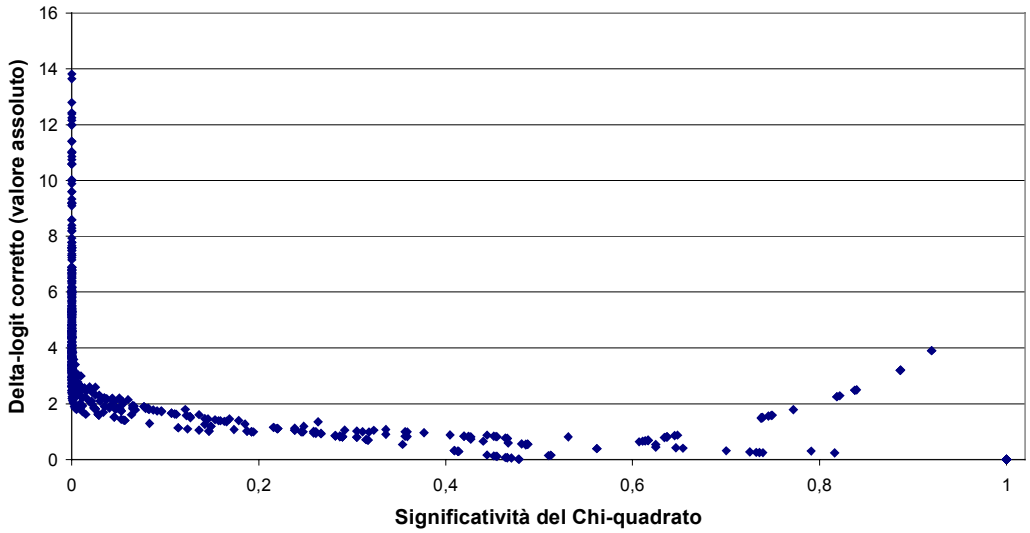
Uno degli obiettivi primari della segmentazione è la ricerca di classificazioni corrispondenti all'ipotesi di discriminazione perfetta[5], o, più verosimilmente, la ricerca di classificazioni in cui almeno alcuni dei gruppi generati abbiano completa omogeneità rispetto alla variabile d'interesse. Questa situazione si riflette nel verificarsi di almeno una frequenza nulla all'interno della Tab. 3 a cui consegue che la statistica $\delta[logit(\pi(Y|X_1, X_0))]$ assume valore infinito. Tuttavia, non sempre una fre-

[5] Si dice *perfetta* la partizione dicotomica che classifica tutte le unità del campione che possiedono l'attributo Y in un sottogruppo e tutte quelle che non lo possiedono nell'altro.

quenza nulla indica una partizione interessante per l'analisi, ma può essere causata dalla bassa frequenza del fenomeno nella popolazione congiuntamente formazione di una classe di modeste dimensioni. Infatti, le tabelle che danno un $\delta[logit(\pi(Y|X_1, X_0))]$ con valore infinito corrispondono a livelli di significatività del test χ^2 che variano tra estremi anche molto lontani (da 0,00 a 0,92; cfr. Fig 1).

Il modo più classico di risolvere i problemi computazionali derivanti da celle vuote in una tabella di frequenze consiste (Agresti, 1986) nell'aggiungere il valore 0,5 alle frequenze nulle nella formula (3). Questa soluzione, pur impedendo che la statistica assuma valore infinito, può causare problemi in una situazione limite, e cioè, quando la caratteristica Y=1 è rara nella popolazione e lo zero corrisponde ad una classe di numerosità esigua: in tal caso, l'aggiunta del valore 0,5 può rovesciare il segno della relazione tra le variabili e indurre a individuare erroneamente la partizione come degna di interesse (Fig. 2).

Figura 2: Relazione tra Chi-quadrato e Delta logit corretto per tabelle con e senza zeri

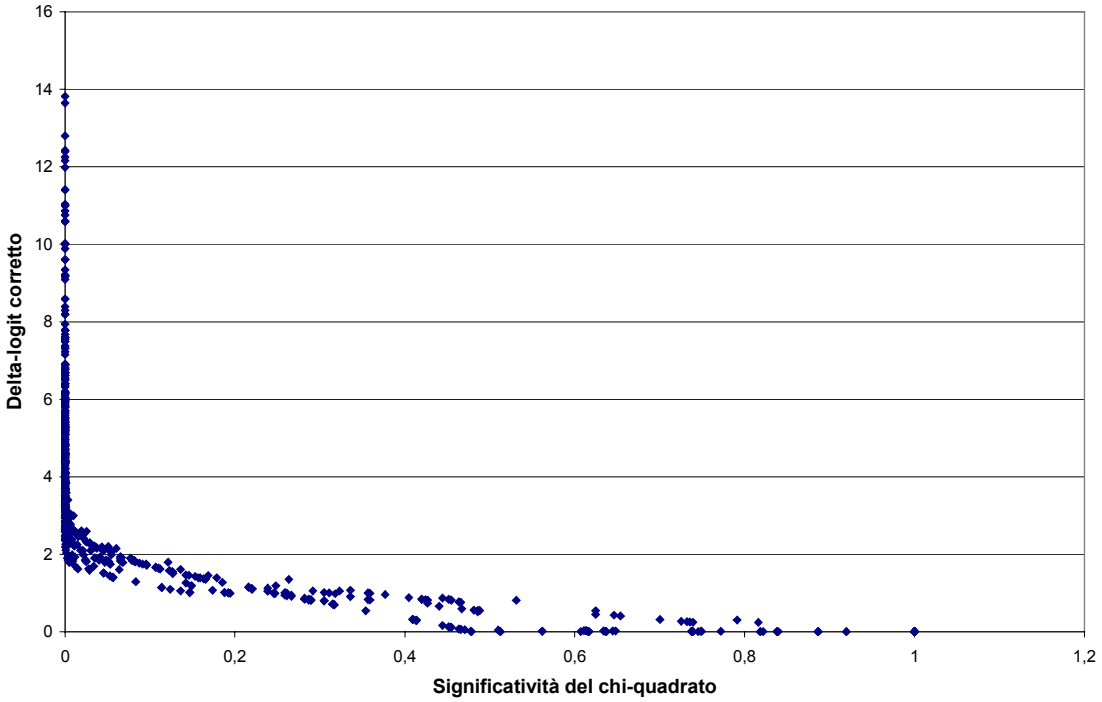


Supponiamo che, nella Tab. 3, sia a=0. Se alla frequenza nulla si sostituisce un valore ε, dato dal più piccolo valore tra 0,5 e un valore funzione di $p(Y=1)$ e della proporzione fra le numerosità delle classi X_0 e X_1,

$$\varepsilon = \min[0{,}5;\ 0.5*c(a+b)/(c+d)], \qquad (5)$$

dove la quantità $c(a+b)/(c+d)$ è il massimo valore di ε che non capovolge la direzione della relazione tra le variabili, la coda anomala è ricondotta su valori molto prossimi allo zero (Fig. 3). Il discorso è analogo quando la cella vuota si trova in un'altra posizione della tabella.

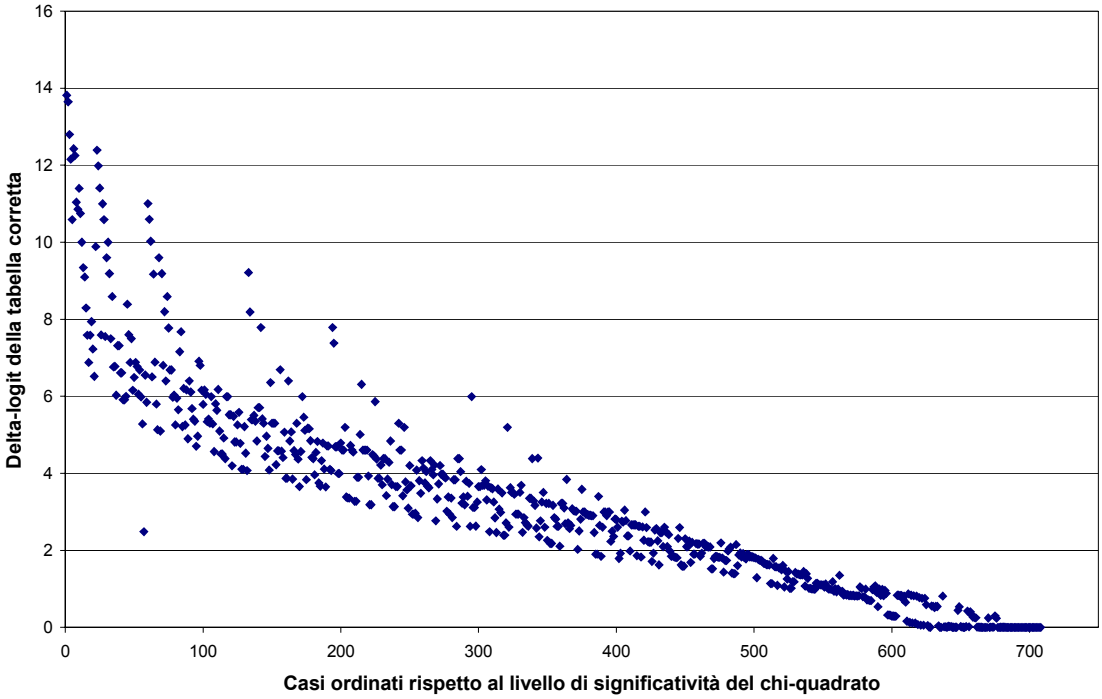
Figura 3: Chi-quadrato e delta-logit con correzione per gli zeri funzione delle frequenze marginali



Le scale su cui sono misurati il criterio basato sui *logit* e quello basato sul χ^2 sono piuttosto diverse. In particolar modo, il criterio basato sui logit differenzia quei casi che il test χ^2 appiattisce su valori prossimi allo zero. Dato che gli specifici valori assunti dai criteri sono irrilevanti per i fini in esame, mentre è rilevante l'ordinamento fra suddivisioni alternative che ne consegue, si ordinano gli insiemi di valori calcolati con il criterio $\delta[logit(\pi(Y|X_1, X_0))]$ rispetto al livello di significatività assunto dal test χ^2. Nella Fig. 4 sono rappresentati graficamente i valori di $\delta[logit(\pi(Y|X_1, X_0))]$ relativi alle situazioni così ordinate.

A fronte di un andamento abbastanza uniforme, e comunque monotòno, si notano alcune "code" di situazioni in cui il criterio $\delta[logit(\pi(Y|X_1, X_0))]$ dà indicazioni molto più positive della corrispondente significatività del test. Si tratta di gruppi di prove che presentano la stessa numerosità totale *n*, ed evidenziano come il metodo basato sui logit privilegi la suddivisione di gruppi, anche di piccole dimensioni, caratterizzate da un rischio relativo elevato, rispetto a suddivisioni di gruppi anche grandi, ma con rischio più modesto. La monotonicità viene completamente recuperata quando si pongono a confronto situazioni relative ad una stessa numerosità totale *n* del nodo.

Figura 4: Valori del Delta-logit con correzione ordinati rispetto al livello di significatività del Chi-quadrato



Il criterio proposto in questa nota si differenzia da analoghi criteri sviluppati in letteratura. A differenza del criterio basato sul test χ^2, ma anche dei criteri utilizzati dal metodo CART (Breiman *et al.*, 1984) e dalla procedura TREE (Venables e Ripley, 1994; Ripley, 1996), il criterio basato sul rapporto crociato è indipendente dalla dimensione del nodo genitore e, pertanto, non privilegia la segmentazione dei nodi di grandi dimensioni. La razionalità del metodo, infatti, non è diretta alla costruzione di un albero bilanciato, quanto alla individuazione di nodi, o sottocampioni, di dimensioni anche minute, caratterizzati da alto rischio di presenza dell'attributo *Y*.

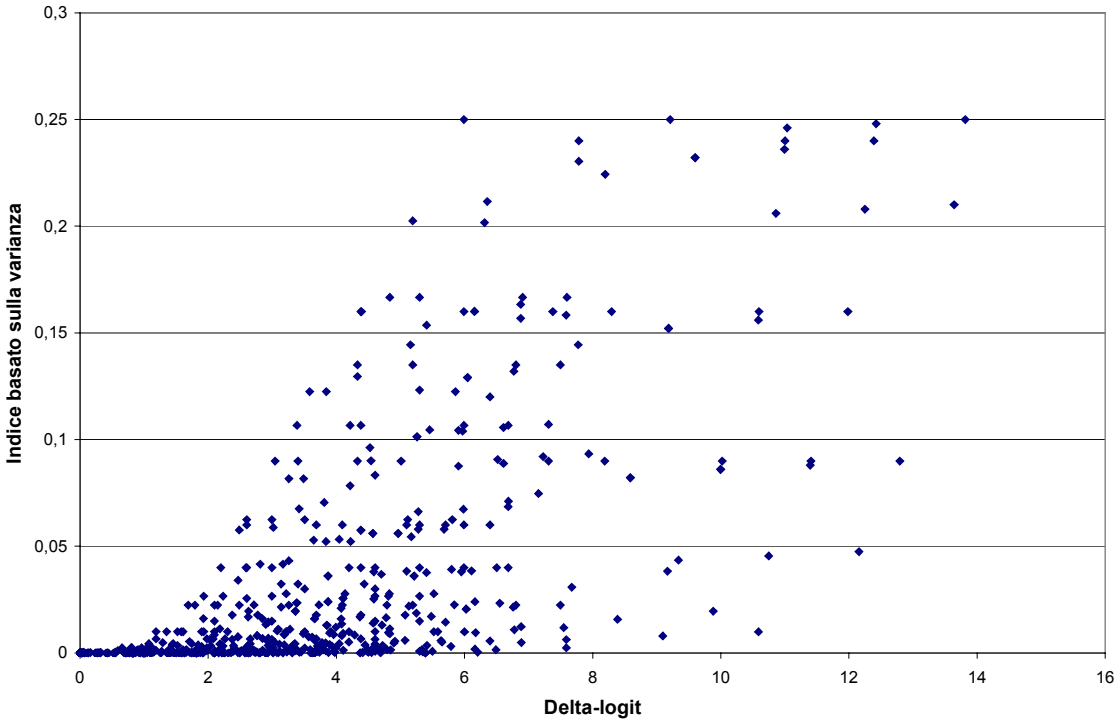
Per garantire robustezza all'analisi, bisogna, comunque, evitare che le frequenze n_i di X_i siano inferiori ad un valore standard, impedendo che l'analisi delle frequenze di *Y* sia svolta entro sottocampioni esigui.

Si possono dare indicazioni generali sul valore al di sotto del quale non si deve proseguire la segmentazione, per esempio, va evitata la formazione di gruppi con numerosità inferiore a 10, oppure si può consentire al ricercatore di specificare tali soglie in modo da tenere conto delle caratteristiche dell'insieme di dati analizzato.

Coerentemente, qualora due suddivisioni concorrenti fornissero lo stesso valore[6] di $\delta[logit(\pi(Y|X_1, X_0))]$, va prediletta la suddivisione effettuata sul nodo di maggiore numerosità, dato anche che ciò corrisponde alla scelta della suddivisione con la più elevata significatività statistica.

Siccome il metodo proposto in questa nota mira solo a massimizzare il rischio relativo in ciascun nodo, le indicazioni fornite da questo metodo e quelle derivanti, tra gli altri, dall'applicazione del metodo basato sulla varianza non sono concordi relativamente a suddivisioni alternative su rami diversi dell'albero (Fig. 5).

Figura 5: Delta-logit e indice basato sulla varianza



Il metodo proposto ricerca "l'ottimalità locale" della funzione obiettivo. Questa peculiarità può indurre a privilegiare - quando si pongano a confronto suddivisioni di nodi diversi - una suddivisione sempre più minuziosa di nodi caratterizzati da frequenze di *Y* molto basse e fortemente diverse nei sottocampioni, invece che procedere alla segmentazione di nodi caratterizzati da un rischio minore (seppure significativo) e con alte frequenze di *Y*. Per risolvere questo problema, conviene porre un

[6] Si tratta di un'eventualità decisamente remota, che potrebbe essere resa più verosimile dalla scelta di considerare un'approssimazione di tale statistica ad una certa cifra decimale, anziché il suo valore esatto.

limite inferiore al rapporto tra la frequenza di *Y* nel nodo in esame e la frequenza marginale di *Y*.

4. L'analisi *Look-ahead*

La suddivisione del campione sulla base ogni volta di un solo predittore candidato esclude dall'analisi le interazioni tra i predittori che non hanno partecipato alla formazione dei gruppi. L'ignorare le interazioni durante il processo di selezione costituisce un rischio per l'analisi sia in termini di capacità esplicativa della variabilità della *Y* da parte dei predittori selezionati sia in relazione alla interpretazione delle combinazioni di modalità che definiscono i gruppi formati.

La procedura *Look-ahead* si suggerisce in quanto studiata per rispondere all'obiettivo di selezionare un predittore alla volta con attenzione sia per gli "effetti principali" su *Y* del predittore stesso, sia per quelli "di interazione" con altri predittori candidati.

Si consideri la possibile suddivisione del campione nei quattro sottoinsiemi definiti dalla combinazione delle modalità dei predittori resi dicotomici[7] X_1 e X_2. Le frequenze campionarie e le proporzioni che possiedono il carattere *Y* nei due sottogruppi sono, rispettivamente, n_{ij} e p_{ij} (i, j=0, 1) (cfr. Tabelle 4 e 5). Si definisce *interazione doppia* la combinazione di modalità di X_1 e X_2 per la quale lo scarto $\hat{\pi}_{ij} - \hat{\pi}_{.j} = p_{ij} - p_{.j}$ risulta maggiore, oppure minore[8], dello scarto $\hat{\pi}_{i.} - \hat{\pi}_{..} = p_{i.} - p_{..}$, o, equivalentemente, quando la frequenza attesa π_{ij} di *Y* all'interno della Tab. 5 è maggiore, o minore, del valore sotto l'ipotesi nulla di assenza d'interazione:

$$E(p_{ij}|H_0\text{: assenza di interazione tra } X_1 \text{ e } X_2) = \pi_{.j} + \pi_{i.} - \pi_{..} . \qquad (6)$$

La procedura che cerca le interazioni doppie è la cosiddetta *Look-ahead* "un passo avanti", la quale consiste nell'espletamento delle seguenti fasi:

a. si esaminano uno alla volta i *q* predittori candidati alla suddivisione per la ricerca del migliore "effetto principale" di ciascuno secondo la procedura illustrata nel Par. 2. $\Delta_i[logit(\pi(Y|X_1, X_0))]$ denota il massimo valore dello scarto tra *logit* inerenti al predittore candidato X_i (i=1, .., q) a cui corrisponde una significatività ($1-\alpha_i$);
b. si incrocia ciascun predittore X_i (i=1, .., q) con uno dei rimanenti predittori candidati X_j ($j \neq i$=1, …, q) e si trova la dicotomizzazione di X_i e di X_j che massimiz-

[7] In realtà, la situazione ipotizzata è una delle $(k-1)(h-1)$ possibilità date dall'incrocio dei due predittori candidati dopo l'ordinamento in base alla frequenza di *y* nelle *k* e *h* categorie, rispettivamente, di x_1 e x_2.

[8] Quando la frequenza p_{ij} supera $p_{.j}+p_{i.}-p_{..}$ si parla di *sinergia* tra le modalità di x_1 e x_2, quando è inferiore si parla di *inibizione*.

Tabella 4. Frequenze campionarie nei sottogruppi che si ottengono

Predittore candidato X_1		Predittore candidato X_2		
		1	*0*	*Totale*
	1	n_{11}	n_{10}	$n_{1.}$
	0	n_{01}	n_{00}	$n_{0.}$
	Totale	$n_{.1}$	$n_{.0}$	$n_{..}$

Tabella 5. Proporzione di unità che, nei sottogruppi, possiedono il carattere *y*

Predittore candidato X_1		Predittore candidato X_2		
		1	*0*	*Totale*
	1	p_{11}	p_{10}	$p_{1.}$
	0	p_{01}	p_{00}	$p_{0.}$
	Totale	$p_{.1}$	$p_{.0}$	$p_{..}$

za uno tra gli scarti tra *logit* calcolabili con le frequenze interne alla Tab. 5 (v. oltre);

$$\ln \frac{\pi_{11}\ (1-\pi_{01})}{(1-\pi_{11})\ \pi_{01}}\ ;\ \ln \frac{\pi_{10}\ (1-\pi_{00})}{(1-\pi_{10})\ \pi_{00}}\ . \qquad (7)$$

Si considerano solo le suddivisioni che generano sottogruppi di numerosità superiore al minimo ammissibile;

c. la suddivisione migliore per ciascun predittore è quella più significativa tra quelle ottenute con la suddivisione "principale" (punto a) e quella con le "interazioni" (punto b). Naturalmente, la suddivisione riguarda solo le modalità della variabile candidata X_i ;
d. Il ramo dell'albero è suddiviso solo se la migliore suddivisione identificata al punto c) è statisticamente significativa al livello prefissato dal ricercatore.

La procedura per la determinazione dei *logit* condizionati alle categorie dicotomiche del predittore X_i (i=1, .., q) segue la falsariga di quella presentata nel Par. 2 per ottimizzare l'aggregazione delle modalità con un singolo predittore osservato su K modalità nominali:

b.1) Si ordinano, dalla più grande alla più piccola[9], le frequenze relative della tabella con K righe e H colonne, dove K è il numero di modalità della variabile X_i e

[9] Se la tabella presenta frequenze uguali, queste possono essere elencate sulla base della numerosità (crescente) delle frequenze campionarie. A parità di numerosità, si possono ordinare sulla base delle modalità riordinate del (primo) predittore candidato X_i . Per esempio, se $p_{1,2}=p_{3,2}=p_{1,3}$, conviene, per motivi pratici, elencare prima $p_{1,2}$, poi $p_{1,3}$, poi $p_{3,2}$ al fine di tenere vicine le modalità di X_i nell'aggregazione.

H quello della X_j . L'ordinamento si presenta come un sequenza di *KH* frequenze, non necessariamente distinte.

b.2) Si riordina la tabella di frequenze relative sulla base dell'ordinamento delle frequenze. Per esempio, se la frequenza più grande è la generica p_{kh} (k=1, .., K; h=1, …, H), la prima modalità della tabella riordinata sarà la k-esima di X_i nel senso delle righe e la h-esima di X_j nel senso delle colonne. La seconda modalità nel senso delle righe e la seconda nel senso delle colonne saranno determinate dai pedici della seconda frequenza nell'ordinamento[10], e così di seguito. Alla fine tutte le frequenze sono collocate nella tabella riordinata.

b.3) Si calcolano, a partire dalla prima posizione, prima sulla prima riga poi sulle successive, i 2(H-1)(K-1) *logit* inerenti alla (progressiva) partizione della tabella in quattro parti secondo le formule (7). I tentativi di partizione della tabella riordinata sono (H-1)(K-1).

b.4) Di ciascuna partizione si salva il *logit* statisticamente più significativo. Questa procedura implica che, in assoluto, la migliore partizione della tabella di K righe e H colonne è una delle (H-1)(K-1) descritte, ossia che nessuna partizione esclusa può essere più efficiente della migliore tra le (H-1)(K-1) tentate.

Per la ricerca di interazioni triple si parte da una tabella tridimensionale di frequenze relative le cui entrate sono le modalità delle variabili X_i , X_j e X_q , di cui la prima è quella della quale si valutano le partizioni. La procedura è una estensione di quella appena presentata per la ricerca delle interazioni doppie:

a. per ciascun predittore, si cerca la migliore suddivisione "principale";
b. si cerca la migliore suddivisione per la determinazione delle "interazioni doppie";
c. si incrocia ciascun predittore X_i (i=1, .., q) con due dei rimanenti predittori candidati X_j e X_m ($m \neq j \neq i$=1, …, q), si riordina la tabella di frequenze sulla base della frequenza della Y e si determina la dicotomizzazione di X_i , X_j e X_m che massimizza uno tra i *logit* calcolabili con le frequenze interne alla Tab. 5:

$$\ln \frac{\pi_{111}\ (1-\pi_{011})}{(1-\pi_{111})\ \pi_{011}};\ \ln \frac{\pi_{101}\ (1-\pi_{001})}{(1-\pi_{101})\ \pi_{001}};\ln \frac{\pi_{110}\ (1-\pi_{010})}{(1-\pi_{110})\ \pi_{010}};\ln \frac{\pi_{100}\ (1-\pi_{000})}{(1-\pi_{100})\ \pi_{000}}. \qquad (8)$$

Si considerano solo le suddivisioni che generano sottogruppi di numerosità superiore al minimo ammissibile;

d. la suddivisione migliore per ciascun predittore candidato X_i è quella più significativa tra quelle ottenute con la suddivisione "principale" (punto a), quella con le "interazioni doppie" (punto b) e quella con le "interazioni triple" (punto c);
e. Il ramo dell'albero è suddiviso solo se la migliore suddivisione identificata al punto c) è statisticamente significativa al livello prefissato dal ricercatore.

[10] Se una di queste è uguale alla precedente, per esempio si tratta della stessa riga, si scrive la frequenza nella seconda colonna della prima riga.

Tabella 6. *Ordinamento delle frequenze per l'analisi Look-ahead*

Modalità di X_i	**Modalità di X_j**	Ordinamento frequenze
$X_{i(1)}$	$X_{j(1)}$	$p_{(1,1)}$
$X_{i(2)}$	$X_{j(2)}$	$p_{(2,2)}$
:	:	:
:	:	:
$X_{i(K)}$	$X_{j(H)}$	$p_{(K,H)}$

Le suddivisioni sono 2^L, dove L è il numero di passi in avanti della procedura: se L=1 si cercano interazioni doppie, se L=2, si cercano interazioni triple, ecc. La determinazione delle interazioni doppie o triple tra predittori non inclusi ad un certo livello dell'analisi di segmentazione annulla il rischio di trascurare importanti predittori dall'analisi. Quantunque non si possano escludere *a priori* interazioni quadruple, o ancora più complesse, tra i predittori non inclusi, il rischio di trascurare predittori importanti dopo il primo o il secondo livello di analisi è praticamente nullo.

Va precisato che la tecnica del *pruning* – la quale espande l'albero fino al massimo livello prefissato e poi lo semplifica eliminando le diramazioni via via meno importanti – non surroga la *Look-ahead*, ma rappresenta una variante della tecnica presentata nel Par. 2. Si potrebbe, invece, congetturare un'espansione dell'albero sulla base della razionalità implicita nella *Look-ahead* e la "potatura" con la procedura *pruning*.

5. Penale per simmetria del dendrogramma

La *penale per simmetria*, se applicata nell'ambito del metodo di analisi della segmentazione sopra presentato, è la frazione F del valore del miglior *logit* cui si è disposti a rinunciare pur di mantenere l'albero simmetrico. L'albero di cui si tratta è simmetrico quando i rami che sono stati formati allo stesso livello sono realizzati sulla base delle stesse modalità della stessa variabile (Fig. 6).

La simmetria dell'albero va nella direzione opposta a quella della massima spiegazione che, per la natura della cose, porta ad alberi molto asimmetrici sia nella forma, sia nelle variabili e nelle modalità utilizzate per le suddivisioni. La simmetria, la quale implica una considerevole semplificazione nel numero e nella struttura delle variabili esplicative, si adotta per intuire dei "modelli" di relazione tra la variabile

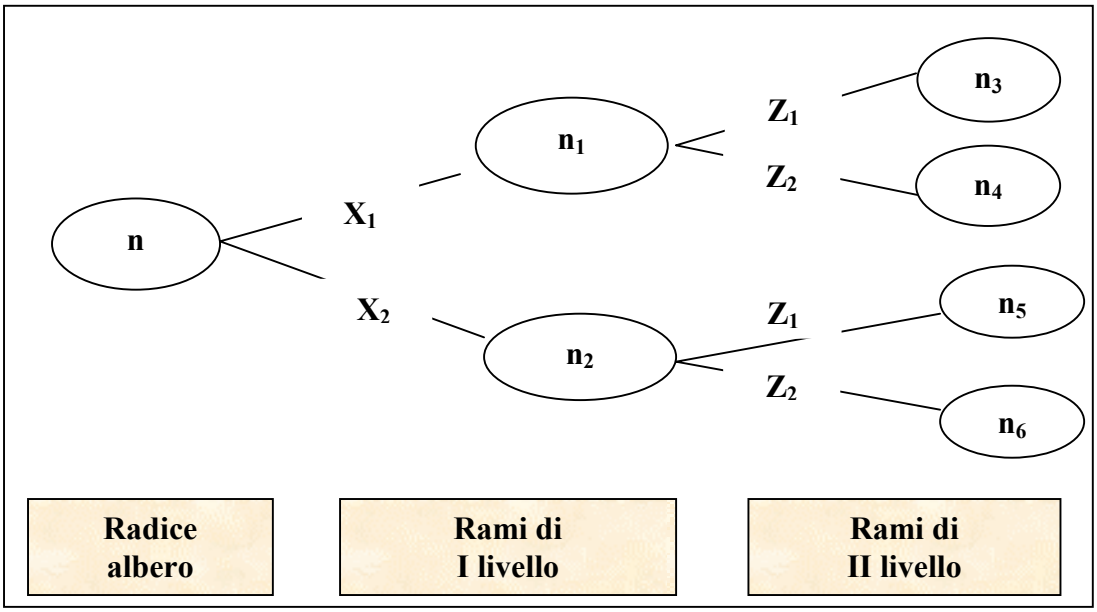


Figura 6. *Esempio di suddivisione simmetrica di un albero diramato su due livelli*

criterio e le variabili predittive[11]. Spesso, infatti, una analisi di questo tipo precede una analisi di classificazione multipla o una analisi della varianza.

La ricerca della simmetria dell'albero si può realizzare seguendo la procedura così schematizzata:

1. si trova la migliore suddivisione ψ_Z* per ogni nodo j (j=1, …, 2^G, dove G è il numero di livelli di segmentazione) e si verifica se è statisticamente significativa. Si ipotizzi, per esempio, che la migliore suddivisione del gruppo (di numerosità) n_1 al secondo livello sia effettuata sulla base delle modalità Z_1 e Z_2 della variabile Z;
2. si cerca la migliore suddivisione parallela ψ_T* sui rami "paralleli" a quelli generati da Z_1 e Z_2 e si verifica se è statisticamente significativa. Si supponga che questa avvenga per effetto della variabile T con modalità T_1 e T_2;
3. per ciascuna suddivisione significativa, si calcola il valore dell'indice ψ_Z e ψ_T per le suddivisioni "simmetriche" da confrontare con quelle ottime su ciascun ramo dell'albero formato. E cioè, la suddivisione trovata al punto 1) si confronta con la suddivisione parallela con la variabile T (modalità T_1 e T_2) e per

[11] La procedura della "penale per simmetria" si può applicare anche per esaminare nel dettaglio, ad ogni livello, o su rami paralleli dell'albero, l'effetto di ciascuna suddivisione. Per esempio, l'esame delle possibili interazioni tra le variabili X e Z della Fig. 6, implica necessariamente che le suddivisioni dei due rami formati con le categorie X_1 e X_2 di X sia svolta sulla base delle stesse modalità Z_1 e Z_2 della Z.

quella trovata al punto 2) si confronta con la suddivisione parallela tramite la variabile Z (modalità Z_1 e Z_2);

4. si effettua una suddivisione simmetrica se l'indice ψ della suddivisione simmetrica a ψ_Z^* dà un rapporto $\psi_Z/\psi_Z^* \geq 1-F$, oppure $\psi_T/\psi_T^* \geq 1-F$. Qualora ambedue le suddivisioni simmetriche siano compatibili con la penale per simmetria, ossia $\psi_Z/\psi_Z^* \geq 1-F$ e anche $\psi_T/\psi_T^* \geq 1-F$, si sceglie quella che dà un rapporto tra indici superiore.

6. Conclusioni propositive

Nella nota si è introdotto un metodo di analisi della segmentazione multivariata di campioni con riferimento ad una variabile dipendente dicotomica trasformata nel logit della frequenza.

Del metodo sono proposti alcuni criteri fondamentali per la ricerca di interazioni. Al confronto con analoghe metodiche di analisi, il criterio si qualifica per la ricerca spinta di rischi, nodo per nodo dell'albero. In questo modo si "scava" a fondo nel campione per la ricerca di gruppi di unità – ossia di combinazioni di modalità dei predittori – caratterizzate da rischi particolarmente elevati.

Il metodo è particolarmente qualificato per la ricerca di fattori di rischio nell'ambito medico-sanitario e in quello naturalistico, e, in generale, in tutte le situazioni di ricerca nelle quali la frequenza del fenomeno in esame è piccola.

Il criterio, come si è detto, pecca di estremismo nella ricerca di rischi e il ricercatore deve porre attenzione nell'analisi dei risultati, soppesando la significatività delle segmentazioni in relazione alla salienza dell'interpretazione. La definizione di una regola che supporti il ricercatore già nella fase della segmentazione può essere un obiettivo dell'automazione della procedura proposta.

La procedura "penale per simmetria" del dendrogramma è, comunque, adeguata non solo per bilanciare la tendenza del criterio ad intensificare la suddivisione di nodi con caratteristiche estreme, ma anche ad identificare modelli interpretativi generalizzabili dei rischi evidenziati. Con questo non si vuol dire che sia necessario un correttivo per l'analisi fondata sul criterio proposto, ma solo che ci possono essere condizioni nelle quali può essere conveniente rinunciare all'ottimalità proposta dall'algoritmo.

La procedura "*Look-ahead*", spingendo l'analisi verso la ricerca di interazioni tra predittori candidati, risolve in parte il problema della preferenza eccessiva per la segmentazione di piccoli gruppi ad alto rischio, ma soprattutto permette l'approfondimento dell'analisi oltre i nodi raggiunti dal metodo ad un certo livello dell'analisi.

Il metodo proposto può essere arricchito con:

a) procedure automatizzate che permettano l'ordinamento dei predittori secondo una concatenazione causale predefinita dal ricercatore, mostrino un dendrogramma autoesplicativo, offrano un'analisi grafica del criterio, presentino una selezione di statistiche qualificanti ciascuna suddivisione operata e un insieme di caratteristiche delle suddivisioni sub-ottimali altrimenti nascoste;
b) uno studio della relazione tra il criterio analitico presentato e l'eventuale forzatura al primo livello dell'analisi di una o più variabili, non necessariamente quantitative, che mirano a definire i sottogruppi di unità entro i quali la relazione tra il logit di $p(Y=1)$ e la variabile forzata è particolarmente significativa. Tra le variabili da forzare si può introdurre anche una o più variabili che qualificano una gerarchia di unità ecologiche che si considerano predittive del rischio;
c) uno studio della suddivisione ternaria del campione rispetto a tutte le opzioni analitiche proposte. La suddivisione ternaria può risolvere problemi di segmentazione sulla base di predittori quantitativi discretizzati, o di predittori ordinali, che non siano in relazione monotòna con la variabile dipendente e che, per questo motivo, possono entrare ripetutamente nella suddivisione dei nodi.

Naturalmente, la convalida e l'arricchimento di significati del metodo possono essere conferiti solo dalle applicazioni su dati reali. Il metodo proposto, infatti, sta alla qualità dei risultati delle applicazioni come il mulino sta alle farine che produce. Queste dipendono dall'adeguatezza del mulino per i tipi di grano che vi entrano. Dipenderebbero anche dal mugnaio, ma questo è proprio un altro tema.

Riferimenti bibliografici

AGRESTI A. (1986) *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New York

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J. (1984) *Classification and Regression Trees*, Wadsworth Inc., Belmont California.

CAPILUPPI C., FABBRIS L. (1998) UNAIDED.2: a PC system for segmentation analysis with a dependent variable on ordinal scale. In: IFCS-98 Secretariat (eds.) *Data Science Classification and Related Methods. VI Conference of the International Federation of Classification Societies (Rome, 21-24 July, 1998)*, SIS-ISTAT, Rome: 48-51

CAPILUPPI C., FABBRIS L., SCARABELLO M. (1999) UNAIDED: a PC system for binary and ternary segmentation analysis. In: VICHI M., OPITZ O. (eds.) *Classification and Data Analysis. Theory and Applications*, Springer, Berlin: 367-374

CELLARD J.C., LABBE B., SAVITSKY G. (1967) Le programme ELISEE. Presentation and application, *Metra*, **3**: 511-519

FABBRIS L. (1997) *Statistica multivariata. Analisi esplorativa dei dati*, McGraw-Hill, Milano

FIELDING A. (1977) Binary segmentation: the Automatic Interaction Detector and related techniques for exploring data structure. In: O'MUIRCHEARTAIGH C.A., PAYNE C. (eds) *The Analysis of Survey Data. Volume 1; Exploring Data Structures*, Wiley, London: 221-257

FISHER W.D. (1958) On grouping for maximum homogeneity, *Journal of the American Statistical Association*, **53**: 789-798

FLACK V.F., CHANG P.C. (1987) Frequency of selection noise variables in subset regression analysis: A simulation study, *American Statistician*, **41**: 84-86

GOODMAN L.A., KRUSKALL W.H. (1954) Measures of association for cross classifications, *Journal of the American Statistical Association*, **49**: 732-764

HOSMER D.W., LEMESHOW S. (1989) *Applied Logistic Regression*, John Wiley & Sons, New York

KASS G.V. (1980) An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, **29**: 119-127

MORGAN J.N., MESSENGER R.C. (1973) *THAID A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables*, Institute for Social Research, Ann Arbor, Michigan

QUINLAN J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufman, San Mateo, California

RIPLEY B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge

SOMERS R.H. (1962) A new asymmetric measure of association of ordinal variables, *American Sociological Review*, **27**: 799-911

SONQUIST J.A., BAKER E.L., MORGAN J.N. (1973) *Searching for Structure*, Institute for Social Research, Ann Arbor, Michigan

VENABLES W.N., RIPLEY B.D. (1994) *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York

Segmentation Analysis with a Logit-transformed Dependent Variable

Summary. *Segmentation analysis is a multivariate statistical method aimed at stepwise partitioning of a sample on which a dependent variable and a set of predictors were observed. If the criterion variable is dichotomous, the conventional segmentation analysis is problematic. Hence, in this paper, we suggest the logit transform of the dependent variable frequencies. The partition of the sample is evaluated, at each analytic step, with reference to the maximum difference between two logit-values of the proportion of Y-variable in the two sub-samples considered for partition. The methodology of analysis is examined in statistical terms also with reference to the look-ahead option for interaction detection, the analysis of a monotone relation between dependent and an ordinal predictor, the so-called "premium for tree symmetry".*

Keywords. *Segmentation analysis; Logit transform; Interaction detection; Monotone relationship; Regression trees; Tree symmetry.*